Predicting Dementia Risk Using Longitudinal Electronic Health Records Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Electronic Health Records (EHR) data are collected as part of routine clinical practice and can be used to train predictive models that estimate and stratify disease risk at the population level. This is particularly valuable for conditions like dementia, where advances in disease-modifying treatments and early interventions have the potential to substantially reduce disease burden. Longitudinal analysis of EHR data can provide valuable insights into dementia risk. In this work, we investigate the utility of statistical learning, graph neural networks, and large language modelling approaches to predict dementia five years before diagnosis using time-course medical history data. We evaluate the performance and utility of five different modelling approaches using data from the UK Biobank (n=9,537) and present a risk stratification model.

1 Introduction

EHR data, routinely collected and stored digitally, capture longitudinal and multimodal information about an individual's health. The ubiquity of EHR platforms across healthcare systems enables population-level and time-aware predictive modelling [1]. Such models can be deployed within existing health information infrastructures, facilitating efficient screening for future health risks without additional data collection or operational costs.

Models of future disease risk could crucially support screening and early detection of prevalent, high-morbidity conditions by increasing the window of opportunity for risk-modifying treatment and early intervention. Dementia is one condition for which early prediction could improve the quality of life of individuals and alleviate pressures on healthcare systems [2, 3]. Although dementia is predominantly diagnosed in late life, dementia-related pathology and distinct comorbidity patterns have been observed up to two decades prior to formal diagnoses [3–6]. In this work, we explore methods capable of modelling the temporal and heterogeneous effects of comorbidities recorded in longitudinal structured EHR data to develop predictive models for dementia risk screening.

2 Background & Related Work

EHR data includes information collected during interactions with various units in healthcare systems, such as diagnostic information, clinical observations and measurements, interventions, and outcomes. Unlike other time-series data, EHR data is episodic and not continuous, meaning the intervals between entries can vary substantially [7]. Probabilistic and statistical learning models often depend on a set of predefined features to develop a predictive model and assume uniform time periods between events. While these models can be highly effective and scalable, the inherent noise and heterogeneity in EHR data make it challenging for models to generalise to wider contexts [8].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Large Language Model (LLM)s have proven effective in modelling EHR data and clinical notes [9, 10]. Several models, pre-trained on EHR data, already exist, such as Med-BERT and Clinical BERT [11, 12]. However, research relying on EHR data often uses pre-processed and curated data sources such as the MIMIC dataset [13]. While these sources are very useful for developing and testing new methods and concepts, they do not reflect the real-world process of working with EHR data directly. In practice, this data can span decades, include repetitive entries and exhibit significant heterogeneity in the temporal dimension [8]. It is crucial to build models capable of capturing this heterogeneity and reflecting the complex effects of these interactions and associated comorbidities, especially in the context of conditions such as dementia.

Conventional approaches to LLMs consider EHR data as a sequence of text over time representing the health trajectory of an individual. However, the temporal patterns and independent and interdependent effects of each comorbidity on the predictive outcome are highly variable. This requires special attention in modelling. Graph Neural Network (GNN)s are a promising alternative for modelling EHR data due to their ability to model complex, non-Euclidean relationships[8]. Message passing, the key learning mechanism of GNNs, enables the propagation of information from each node to its neighbours, allowing the model to learn interactions across components of EHR data [14]. Consequently, interactions between components of EHR data are preserved. Temporal edge aggregation provides a mechanism for representing temporal relationships between nodes [15, 16]. GNNs have previously been used to model temporal dependencies in EHR data [17], but they have not been used for risk prediction using temporal patterns of comorbidities in complex conditions such as dementia.

3 Methods

3.1 Data & cohort selection

Data source All analyses were conducted using data from the UK Biobank, an ongoing study of more than 500,000 people living in the United Kingdom (UK). Up to 30 years of historical EHR data is available for each participant, with each diagnosis coded according to the ICD-10 system and accompanied by the date of diagnosis.

Cohort selection Dementia and control cohorts were defined as follows: Participants were included in the dementia cohort if their historical EHR data contained an ICD-10 code beginning with F00 or G30 (Alzheimer's disease (AD)), or F01(Vascular dementia (VD)), corresponding to the two most common forms of dementia. The control cohort was age- and sex-matched to the dementia cohort based on the date when each participant was diagnosed with dementia, and included participants with no record of a dementia diagnosis. Excluded codes are listed in Appendix A1. The final cohort consisted of 9,537 patients.

Features Participants' diagnosis histories at five years prior to their first dementia diagnosis or index date, for those in the control cohort, were used in this study. Two features were generated for each diagnosis in a participant's EHR: a binary feature to indicate presence, and a continuous measure of the time (in days) between the diagnosis date and the date five years before a dementia diagnosis. Additionally, participants' age, sex, and polygenic risk score (PRS) for AD were included.

3.2 Model architecture & training

Two baseline and three deep learning models were assessed for five-year dementia risk classification.

Baseline To capture the relationship between presence and time features, two tree-based model architectures were assessed for baseline model training: Random Forests and XGBoost. Hyperparameters were tuned using a Bayesian search with 5-fold cross-validation. Each model pipeline included a SelectFromModel feature selection step to select the top 25% of features ranked by importance from the model being optimised.

LLM: BioClinical BERT BioClinical BERT was used to capture the medical context and temporal relationship between a patient's diagnoses. A single string was generated for each patient containing their sex, age, and AD PRS followed by each diagnosis and the corresponding time since they were diagnosed, e.g. *"sex: male, age: 65, polygenic risk score: 0.125, hernia 6 months ago"*. Strings were

tokenised to the 512-token maximum of BioClinical BERT and fine-tuned for sequence classification to predict dementia.

Multimodal modelling: BioClinical BERT + MLP Next, we assessed a multimodal classifier that encoded diagnoses and temporal components with the same text pipeline as the BioClinical BERT-only approach. Structured data (age, sex, PRS) were modelled separately using a two-layer Multilayer perceptron (MLP) with ReLU and dropout. The MLP embedding was concatenated with the BERT [CLS] embedding (i.e. the final hidden state of the classification token), and the resulting vector was passed to a classification head to produce two logits, one for each class.

GNN architecture Each patient was represented by a complete bipartite star graph ($K_{1,k}$) where the centre patient node was connected to each of the patient's diagnoses. The features of the centre node were the patient's sex, age, and PRS score, and each diagnosis node was only connected to the centre node. The node features of the diagnosis nodes were initially stored as placeholder vectors, and true diagnosis feature representations were generated during training using precomputed BioClinical BERT embeddings (dimension 768) or randomly initialised embeddings (dimension determined via hyperparameter tuning). Edge attributes were used to model the temporal component of each diagnosis, consisting of up to three temporal features: years since diagnosis, natural logarithm of days ($\log_e(1+x)$) which highlights smaller time differences, and temporal decay, calculated by taking the exponential decay of days with a decay constant of 180 (Equation 3, Appendix). GINEConv (Graph Isomorphism Network with Edge features) was chosen as the GNN architecture, due to its ability to directly incorporate edge attributes ($\mathbf{e}_{j,i}$) during training (see Equation 1) [18, 19].

$$\mathbf{x}'_i = h_\Theta \left((1+\epsilon) \cdot \mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} \mathrm{ReLU}(\mathbf{x}_j + \mathbf{e}_{j,i}) \right) \tag{1}$$

Patient and diagnosis node embeddings were combined to run through two GINE layers. The resulting GINE embeddings were pooled, either by mean, addition, or max (determined via hyperparameter tuning), to receive a single graph-level representation. Two alternative classification head designs were evaluated: (1) a linear layer to map the pooled representation to the output, and (2) an MLP with a hidden layer, ReLU activation, and dropout, followed by the final output layer for classification of the two labels.

Training For all models, data was split into training and testing sets at a ratio of 80:20. All models were evaluated using sensitivity, specificity, F1 score, and area under the receiver operating characteristic curve (AUCROC). The best overall model was selected using Youden's J index (2).

$$J = \text{Sensitivity} + \text{Specificity} - 1 \tag{2}$$

The three deep learning methods all used weighted Adam for optimisation and cross-entropy as the loss function. The first two models were trained for up to 20 epochs, whereas the GNN-based models were trained for up to 50 epochs. In all cases, early stopping was implemented, and validation loss was monitored to avoid overfitting. Extensive hyperparameter tuning was conducted using Bayesian optimisation to determine multiple hyperparameter values (listed in Appendix Table A1).

Risk stratification Predictions were stratified into three groups based on their predicted class probabilities. Stratification allows for focused prediction of those at greater or lesser risk and facilitates streamlined clinical decision-making. Youden's J was used to determine the threshold that maximises J-index for the dementia class. Predictions for the stratified groups were reevaluated.

Explainability SHapley Additive exPlanation (SHAP) values were calculated for the baseline model, and two GNN explainers, Gradient and Back Propagation Explainer, were used for GNN explainability [20]. The code was adapted from GraphXAI's GitHub [21]. The overall sum and mean of the Explainer importance scores were used to identify the most influential diagnoses for dementia prediction, and mean absolute SHAP values were determined for baseline model features.

4 Results

A GNN with a linear classification head, using trainable, randomly initialised vectors (GNN+RV) achieved the best overall performance. The hyperparameters selected during training are in Ap-

pendix Section A3. Performance on all models is outlined in Table 1. Following risk stratification, performance on both the XGBoost and GNN+RV models improved, with XGBoost marginally outperforming GNN+RV on the J-index. Stratification thresholds can be seen in Appendix Table A2.

Table 1: Five-year dementia risk model performance. RV = random initialised diagnosis vectors, BioClinical = BioClinical BERT initialised diagnosis vectors

Models before risk stratification	**F1**	**Sensitivity**	**Specificity**	**J**	**AUCROC**
GNN+MLP+RV	0.709	0.696	0.735	0.428	0.780
GNN+RV	**0.716**	**0.761**	**0.673**	**0.434**	**0.777**
GNN+BioClinical	0.709	0.746	0.672	0.415	0.775
GNN+MLP+BioClinical	0.713	0.753	0.673	0.426	0.770
BioClinical BERT pure	0.707	0.726	0.670	0.396	0.767
Multimodal BioClinical BERT+MLP	0.704	0.694	0.714	0.408	0.776
XGBoost	0.708	0.705	0.711	0.416	0.773
Models after risk stratification	**F1**	**Sensitivity**	**Specificity**	**J**	**AUCROC**
GNN + RV	**0.816**	**0.834**	**0.758**	**0.592**	**0.838**
XGBoost	0.810	0.821	0.780	0.601	0.844

Explainability methods were applied to the XGBoost and GNN+RV models. Both GNN Explainers identified hypertension, arthropathies, and diseases of oesophagus, as the top three most influential nodes in dementia risk classification. The Gradient Explainer identified "general symptoms and signs", "symptoms and signs involving the digestive system and abdomen", and "symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" to be the next most predictive. Back Propagation Explainer similarly identified these categories and additionally ranked "factors influencing health status and contact with health services" among the top ten most influential nodes. Age and PRS had the strongest influence on XGBoost predictions, followed by hypertension presence, and time since "factors influencing health status and contact with health services" and diabetes diagnoses. The full table of importance scores can be found in Appendix Table A3.

5 Discussion

Across all models, GNNs achieved higher performance. In dementia, both the duration of comorbidities and the interactions between them influence risk [3]. GNNs represent both of these factors by (1) encoding the temporal influence of comorbidity duration in the graph edges, and (2) modelling the influence of co-existing diagnoses on dementia risk through message passing [14]. Unlike other deep learning methods, they offer explainability, a crucial feature of clinical predictive models [21]. GNNs are therefore well-suited to represent the complex temporal interactions between pre-existing conditions while offering insight into the features driving model decisions.

Our models were designed to classify dementia risk while meeting clinically relevant performance metrics with a particular focus on deriving insights into how comorbidities and their temporal patterns contribute to dementia risk. Beyond prediction, our GNNs provide a framework that can handle incomplete data and simulate how changes in comorbidity duration or profile may alter risk levels.

We chose a five-year prediction time-frame because it offers a "window of opportunity" for disease-modifying interventions [6, 22], while balancing the degree of uncertainty in risk prediction [23]. The implemented risk stratification system allows for streamlined clinical decision-making and follow-up by categorising patients into high, moderate, and low risk categories.

Our findings show that the GNN and XGBoost models can successfully assess dementia risk five years before diagnosis using only the time-course ICD10 codes, PRS, age, and sex. Due to the standardised nature of ICD-10 codes, models developed using this framework can be used in a wide variety of clinical settings. These findings demonstrate the potential for GNN-based models to be used for time-course-aware, EHR-based dementia risk screening.

References

[1] Sonia Akter et al. "Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias". In: *The Journal of Prevention of Alzheimer's Disease* (2025), p. 100169.

[2] GBD 2019 Dementia Forecasting Collaborators. "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019". In: *Lancet Public Health* 7.2 (2022), e105–e125. DOI: 10.1016/S2468-2667(21)00249-8.

[3] Gill Livingston et al. "Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission". In: *The Lancet* 404.10452 (2024), pp. 572–628. DOI: 10.1016/S0140-6736(24)01296-0.

[4] Michael S. Rafii and Paul S. Aisen. "Detection and treatment of Alzheimer's disease in its preclinical stage". In: *Nature Aging* 3.5 (2023), pp. 520–531. DOI: 10.1038/s43587-023-00410-4.

[5] Richard J. Caselli et al. "Neuropsychological decline up to 20 years before incident mild cognitive impairment". In: *Alzheimer's Dementia* 16.3 (2020), pp. 512–523. DOI: 10.1016/j.jalz.2019.09.085.

[6] Chloe Walsh et al. "Chronological Mapping of Comorbidities in Alzheimer's Disease and Vascular Dementia". In: *medRxiv* (2025). DOI: 10.1101/2025.04.28.25326575. URL: https://www.medrxiv.org/content/early/2025/04/29/2025.04.28.25326575.

[7] Vinod Kumar Chauhan et al. "Continuous patient state attention model for addressing irregularity in electronic health records9". In: *BMC Medical Informatics and Decision Making* 24.17 (2024). DOI: https://doi.org/10.1186/s12911-024-02514-2.

[8] Zheng Liu et al. "Heterogeneous similarity graph neural network on electronic health records". In: *2020 IEEE international conference on big data (big data)*. IEEE. 2020, pp. 1196–1205.

[9] Jiheum Park et al. "Enhancing EHR-based pancreatic cancer prediction with LLM-derived embeddings". In: *npj Digital Medicine* 8.1 (2025), p. 465.

[10] Hanan Alghamdi and Abeer Mostafa. "Advancing EHR analysis: Predictive medication modeling using LLMs". In: *Information Systems* 131 (2025), p. 102528.

[11] Laila Rasmy et al. "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *NPJ digital medicine* 4.1 (2021), p. 86.

[12] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://www.aclweb.org/anthology/W19-1909.

[13] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (Jan. 2023), p. 1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. URL: https://doi.org/10.1038/s41597-022-01899-x.

[14] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

[15] Antonio Longa et al. "Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities". In: *arXiv preprint arXiv:2302.01018* (2023).

[16] Siyue Xie et al. "GTEA: Inductive representation learning on temporal interaction graphs via temporal edge aggregation". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2023, pp. 28–39.

[17] Jiayuan Chen et al. "Predictive modeling with temporal graphical representation on electronic health records". In: *IJCAI: proceedings of the conference*. Vol. 2024. 2024, p. 5763.

[18] Keyulu Xu et al. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).

[19] Weihua Hu et al. "Strategies for pre-training graph neural networks". In: *arXiv preprint arXiv:1905.12265* (2019).

[20] Antonio Longa et al. "Explaining the explainers in graph neural networks: a comparative study". In: *ACM Computing Surveys* 57.5 (2025), pp. 1–37.

[21] Chirag Agarwal et al. "Evaluating Explainability for Graph Neural Networks". In: *Scientific Data* 10.144 (2023). URL: https://www.nature.com/articles/s41597-023-01974-x.

[22] Anne Corbett et al. "Cognitive decline in older adults in the UK during and after the COVID-19 pandemic: a longitudinal analysis of PROTECT study data". In: *The Lancet Healthy Longevity* 4.11 (Nov. 2023), e591–e599. ISSN: 2666-7568. DOI: 10.1016/S2666-7568(23)00187-3.

[23] Laura D Baker et al. "Study design and methods: US study to protect brain health through lifestyle intervention to reduce risk (US POINTER)". In: *Alzheimer's & Dementia* 20.2 (2024), pp. 769–782.

A1 Exclusion criteria

Participants with ICD-10 diagnoses beginning with F02, F03, F04, F05, F06.7, G31, G32, or Q90 were excluded from both cohorts.

A2 Graph Neural Network Architecture

Formula 3 was used to apply a temporal decay to the time since diagnosis. Diagnoses that happened more recently are assigned a number closer to one.

$$f(\text{days}) = \exp\left(-\frac{\text{days}}{180}\right) \tag{3}$$

A3 Hyperparameter tuning search space

Table A1: Hyperparameter search space

Hyperparameter	**Search space**	**Applicable models**
Learning rate	log-uniform $[3 \times 10^{-6},\ 3 \times 10^{-3}]$	All neural models
Trainable embeddings	$[True, False]$	All GNN-based models
Learning rate for trainable embeddings	log-uniform $[3 \times 10^{-4},\ 3 \times 10^{-3}]$	GNN models (trainable emb.)
Random initialised embeddings dimension	$[96, 128, 192, 256]$	GNN models (random. emb.)
Batch size	$\{8, 16, 32\}$	All neural models
Dropout	uniform $[0.0,\ 0.5]$	BERT/MLP heads
Hidden dim (MLP)	$\{32, 64, 96, 128\}$	MLP components
Weight decay	log-uniform $[1 \times 10^{-6},\ 1 \times 10^{-3}]$	All neural models
Pooling method	$[mean, add, max]$	All GNN-based models
Min child weight	$\{1, 2, \ldots, 10\}$	XGBoost
gamma	log-uniform $[0.1,\ 5]$	XGBoost
Subsample ratio	uniform $[0.4,\ 1.0]$	XGBoost
Column subsampling ratio	uniform $[0.4,\ 1.0]$	XGBoost
Max depth	$\{3, 4, \ldots, 10\}$	XGBoost

RV: randomly initialised diagnosis vectors

The hyperparameter search space values for all models are provided in Table A1. The GNN model with trainable, randomly initialised vectors had the best overall performance recorded at epoch 14. The optimal hyperparameter values were as follows: pool: mean, hidden dimension (MLP): 128, learning rate for trainable embeddings: 0.00016, random initialised embedding dimension: 96, dropout: 0.474, learning rate: 0.00016, weight decay: 0.00016. The hyperparameters for the best performing XGBoost model were: minimum child rate: 1, gamma: 4.999, subsample ratio: 1.0, column subsampling ratio: 0.4 and maximum depth: 10.

A4 Stratification thresholds

The thresholds for the stratification of the best performing model (GNN+RV) and the XGBoost can be seen in A2. These were determined using Youden's J index (equation 2).

Table A2: The threshold values for the stratification method, determined using Youden's J index.

Model	**Green (Control)**	**Amber (Uncertain)**	**Red (Dementia)**
GNN+RV	[0, 0.305)	[0.305, 0.697)	[0.697, 1]
XGBoost	[0, 0.315)	[0.315, 0.731)	[0.731, 1]

A5 GNN Explainers

Gradient Explainer uses backpropagation to calculate importance scores. The explainer uses the magnitude of a node's final derivative to assign importance. The larger the magnitude, the more important the node, due to the removal of the node having a larger impact on the predictions [20].

BackProp Explainer follows the same approach as Gradient Explainer, but differs by not considering negative gradients, regarding them as noise [20]. Consequently, BackProp Explainer only focuses on nodes contributing in a positive manner to predictions.

A6 Explainability results

Table A3 provides the numerical results for the different explainability methods. The GNN Explainers are evaluated using both the overall sum of importance for a node and the mean, providing an overview of the influence at both a global (sum across all occurrences) and per-occurrence (mean) level.

Table A3: Explainability results. RV = randomly initialised vectors

Gradient Explainer applied to GNN+RV		
Top 5 features	**Sum of scores**	**Mean of scores**
Hypertensive diseases	147.626	0.214
Arthropathies	113.739	0.228
Diseases of oesophagus, stomach and duodenum	78.090	0.174
General symptoms and signs	67.089	0.200
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	53.574	0.147
Guided Backpropagation Explainer applied to GNN+RV		
Top 5 features	**Sum of scores**	**Mean of scores**
Hypertensive diseases	74.295	0.108
Arthropathies	71.554	0.144
Diseases of oesophagus, stomach and duodenum	40.193	0.090
Factors influencing health status and contact with health services	33.037	0.774
General symptoms and signs	32.244	0.096
Mean absolute SHAP values for XGBoost		
Top 7 features	**SHAP**	
PRS	0.547	
Age	0.470	
Hypertensive diseases: present	0.090	
Factors influencing health status and contact with health services: time	0.088	
Diabetes mellitus: time	0.059	
Diabetes mellitus: present	0.047	
Disorders of lens: present	0.045	

A7 Fairness analysis

The performance of each model was assessed across sexes to investigate any key differences in model fairness. Metrics are outlined in Table A4

Table A4: Five-year dementia risk predictive model performance, by sex

Model	**Sex**	**F1**	**Sensitivity**	**Specificity**	**AUCROC**
XGBoost	F	0.699	0.688	0.711	0.773
	M	0.709	0.704	0.714	0.770
GNN+RV	F	**0.702**	**0.761**	**0.671**	**0.777**
	M	**0.702**	**0.761**	**0.671**	**0.777**

A8 Implementation Details

The models were run with Python 3.11.13. The following packages were used: Pandas (v. 2.3.1), NumPy (v. 1.26.4), PyTorch (v. 2.7.0), PyTorch Geometric (v. 2.6.1), transformers (v. 4.54.1) and scikit learn (v. 1.7.1). The deep learning models were run on one NVIDIA A100 80 GB GPU. Weights & Biases (WandB) was used to perform hyperparameter tuning.

A9 Acknowledgements

This research has been conducted using the UK Biobank resources under application number [number to be included post review]. We are incredibly grateful to the UK Biobank participants for their contributions to this research and the broader research community.